Australian Government
Department of Defence
Defence Science and
Technology Organisation

A Tracker Assessment Tool for Comparing Tracker Performance

S. B. Colegrove, S.J. Davey and B. Cheung

**Intelligence, Surveillance and Reconnaissance Division
Information Sciences Laboratory**

DSTO–TR–1694

ABSTRACT

This technical report describes a method for assessing the performance of automatic tracking systems against various metrics. These metrics are grouped in the categories of track establishment, track maintenance, track error and false tracks. A desired performance level is defined for each metric, and the probability that a particular track will meet this performance level is empirically approximated from archived sensor data. The individual metric probabilities are combined using a weighted average to produce an overall tracker rating. An important feature of the approach is that absolute truth is assumed to be unavailable, and output of the trackers under test is instead compared with *incomplete truth.* The incomplete truth is produced by a manual inspection of the data and through editing the output of the best available tracking algorithm. This report also discusses an implementation of the assessment method for Over-the-Horizon Radar. This software package is referred to as the Tracker Assessment Tool, and is used to compare four alternative tracking algorithms.

Published by

DSTO Information Sciences Laboratory
PO Box 1500
Edinburgh, South Australia, Australia 5111

Telephone: (08) 8259 5555
Facsimile: (08) 8259 6567

AR No. AR 013-351
March 2005

A Tracker Assessment Tool for Comparing Tracker Performance

EXECUTIVE SUMMARY

This report describes a method that was developed for assessing the performance of automatic tracking systems using recorded data as well as simulated data. An ability to assess tracking performance is useful for testing and optimizing a new tracking filter, comparing alternative tracking systems, and evaluating contract specification compliance. A secondary role is to use tracking performance to test and optimize signal processing and detection routines.

The initial ideas to solve the problem of measuring tracking performance came from research and development of automatic detection and tracking systems for the Jindalee Over-the-Horizon Radar (OTHR). The first approach was formulated in 1991 and this lead to the start of work on a software package for tracker performance assessment, referred to as the Tracker Assessment Tool (TAT). At the same time, a panel of international specialists on Radar Data Processing (KTP-2) was investigating the characterisation of tracker performance for Microwave radars. The KTP-2 panel operated under Sub Group K (Radar) within The Technical Cooperation Program (TTCP). At the Panel's 1994 meeting, a brainstorm session that involved Colegrove resulted in a list of metrics to describe tracking performance.

Further development of the assessment method and the TAT has occurred as a result of lessons learned through using it. This paper defines the approach for measuring tracker performance using the TAT. It includes a description of the performance metrics which extend and refine the KTP-2 metrics, and provides a method for approximately quantifying the precision of the results. In addition, the TAT has been extended to do assessment over an arbitrary number of data sets. The procedures for obtaining truth track data from experimental data are addressed with the solution that has evolved through extensive use of the TAT. To illustrate the assessment process, the results are given from the comparison of four competing tracking systems for the Jindalee OTHR.

Authors

Samuel B Colegrove
Intelligence Surveillance and Reconnaissance Division

Colegrove received his BE and PhD in Electrical Engineering at the University of Queensland in 1969 and 1974 respectively. He joined the Defence Science and Technology Organisation in 1973. Shortly after starting, he transferred to work on Over-the-Horizon Radar (OTHR) in 1974 where he has been to this day. He was a member of the design team for the Jindalee Stage B OTHR. He participated in the implementation of signal processing, displays and an automatic tracking system for Jindalee. From 1979 to 1994 Dr Colegrove was the Australian National Leader of TTCP Technical Panel KTP-2 whose charter was on advancements in automatic tracking systems. Dr Colegrove has also participated in tender evaluation and performance evaluation of US and Australian OTHR tracking systems. Dr Colegrove's research activities, prior to leaving DSTO in December 2004, were into advanced automatic tracking systems.

Samuel J Davey
Intelligence Surveillance and Reconnaissance Division

Samuel Davey received the Bachelor of Engineering, Master of Mathematical Science, and PhD degrees from the University of Adelaide in 1996, 1999 and 2003 respectively. In 1995 he joined the Defence Science and Technology Organisation, where he has worked on tracking system performance assessment, design of real time multi-target tracking algorithms, automatic track initiation, and multi-sensor fusion. His current research interests include decentralised data fusion, multi-sensor fusion performance analysis, and probabilistic multi-hypothesis tracking.

Brian Cheung
Intelligence Surveillance and Reconnaissance Division

Brian Cheung graduated from the University of South Australia in 2001 with a Bachelor of Engineering degree. Since 2001 he has been employed as a Professional Officer at the Defence Science and Technology Organisation (DSTO) where he has worked in the fields of advanced automatic tracking and tracking performance assessment. His background is in computer systems engineering and his research interests are in signal processing and advanced automatic tracking systems. He is currently studying towards his Masters degree in Signal and Information Processing at the University of Adelaide.

Contents

Appendices

Figures

Tables

1 Introduction

An ability to assess the performance of a tracking system is primarily useful for: (1) new tracking filter testing and optimization, (2) comparison of alternative tracking systems, and (3) evaluation of contract compliance. A secondary role is to use tracking performance to test and optimize signal processing and detection routines. In general, algorithm assessment requires the definition of two things: the ideal algorithm output (*truth*) and comparison rules for quantifying the difference between the achieved output and the ideal output. The definition of truth is trivial when simulated data are used; however, algorithm performance on real sensor data is more important, since this reflects system performance. Determining the truth for real sensor data is often very difficult. For multi-target problems, the assessment process must also associate individual output tracks with truth tracks, the result of which may be ambiguous. The second requirement is the definition of comparison rules. For tracking, there are many different aspects of performance that need to be compared, and each of these needs to be considered to obtain an overall measure of performance. Each different comparison metric will generally lead to a distribution of results. For example, the time to establish a track on a new target is a random quantity with a different distribution for each tracker. Such data may need to be summarised in order to decide which algorithm provides superior performance. In many cases, the aim of assessment is to make a hard decision about which algorithm is preferred for a particular situation. This is made difficult by the multiplicity of comparison metrics; where one algorithm may excel in some metrics, another may be preferred under other metrics. In order to make a hard decision, it is necessary to distil these possibly conflicting and numerous metrics into a single performance measure that summarises all of the results and objectively trades off good performance in one area against poor performance in another. Each of these difficulties is addressed by the performance assessment method described in this report, and implemented in the Tracker Assessment Tool (TAT).

In early work on tracker assessment, [Castella 1986] developed an automatic procedure for assessing tracking systems by fitting a least squares quadratic curve to a track's input data for each one minute interval. If the weighted sum of the squared deviations of the radar measurements from this curve exceeds a threshold for 50% of intervals, the track is declared false. This procedure avoids manual procedures for defining valid tracks and analyses only those tracks produced by the tracking system. Castella's method characterises the false track and accuracy performance of a tracker, but does not give other important aspects such as the establishment and maintenance of a track. More recently, procedures have been proposed for the performance evaluation of single and distributed sensor trackers in [Rothrock & Drummond 2000] and [Mason & O'Kane 1992]. In [Rothrock & Drummond 2000], the tracking system output is compared with the true object state. The comparison uses a unique assignment between each track and each object. That is, no object is assigned to more than one track. Other tracks that are near to an object which is already associated are defined as redundant tracks. A set of metrics are defined which cover track output performance as well as network track number commonality and tracker processing load. Each metric is measured in a different unit as time proceeds. No attempt is made to give an overall measure of performance.

The alternative approach in [Mason & O'Kane 1992] classifies the tracks as 'clean', 'wild' or 'other'. These are further subdivided to identify cases such as redundant and dropped tracks. This leads to a total of 17 track categories. An expression is given for an overall performance which is based on the number of tracks in each category. These numbers require an operator to manually determine the track categories. The authors do comment that this process could be automated. Because this procedure deals only with categorizing the track output, key parameters such as initiation delay and track overshoot cannot be measured.

At DSTO, the research and development of automatic detection and tracking systems for the Jindalee Over-the-Horizon Radar (OTHR) motivated investigation into developing a method to measure tracking performance. An initial approach was formulated in 1991 [Colegrove & Mabbs 1991] and this lead to the creation of a software package, referred to as the Tracker Assessment Tool (TAT) [Mabbs 1993]. At the same time, a panel of international specialists on Radar Data Processing (KTP-2) investigated the characterisation of tracker performance for Microwave radars. The KTP-2 panel operated under Sub Group K (Radar) within The Technical Cooperation Program (TTCP). At the Panel's 1994 meeting, a brainstorm session that involved Colegrove resulted in a list of 15 metrics [1994] to describe tracking performance.

Since that time, further development of the TAT has occurred from lessons gained through using it [Colegrove et al. 1996]. This paper defines the approach for measuring tracker performance by the TAT. It includes a description of the metrics which extend and refine the KTP-2 metrics. The main extension to that described in [Colegrove et al. 1996] is the method for assessing performance over an arbitrary number of data sets [Colegrove et al. 2003]. Also the procedures for obtaining truth track data from experimental data are addressed with the solution that has evolved through extensive use of the TAT. To illustrate the assessment process, an example is given of the displays and output from the comparison of Australian OTHR trackers.

2 Tracker Assessment

A valid source of truth tracks is essential for tracker assessment. Simulations allow truth to be derived from the simulated track parameters, while recorded data can use secondary radar tracks and GPS flight logs. In some cases, particularly with OTHR, it is impossible to obtain track reports for all objects. Because of this, the following procedure can use either known truth data or truth data derived from recorded data. Since the latter is potentially error prone, the truth is referred to as *incomplete* truth. That is, there is not complete object track knowledge.

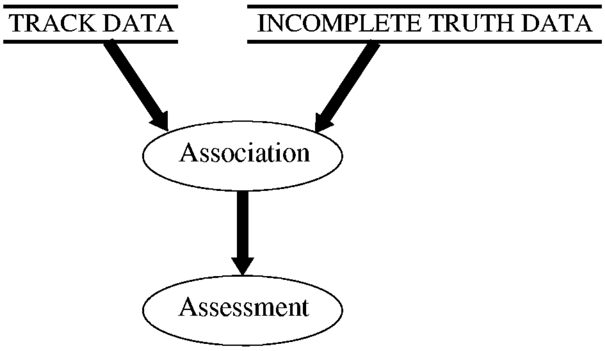


Figure 1: Tracker Assessment Functions

Figure 1 shows the data sources and basic processes required for tracker assessment. Here the Track Data are abbreviated as TK tracks and the Incomplete Truth Data are abbreviated IT tracks. TK tracks are from the tracking system under test and IT tracks are the best estimates of tracks for the real objects. The *association* process correlates TK and IT data. The final process of *assessment* calculates the metric ratings from the associations and the tracks in the TK and IT data. The TAT processes are based on a sequential progression through time-aligned TK and IT data with association performed at each time step 'k'. The primary intent of the TAT and its metrics is to assess the real-time output of a tracking system.

If a tracking system alters its history as time advances, the measured performance is dependent on what track history estimates are stored in the TK data at each time step. When the TK data contains the stabilised history estimates, it will give different results to TK data that contains the real-time track estimates. So for comparing tracking systems with differences such as track history modification, the TK data must be written to minimise any bias in the results. For the case of tracking systems that don't alter history, the TK track data is written on promotion of each track for display to an operator. It ceases to be written when the track is deleted.

The tracker should log all tracks in the TK data, not only those that would be reported to the user. In addition to state estimates, the track confidence (quality) of each track is recorded. This is the internal measure used by the track maintenance functions for deciding when to terminate tracks and when to report them to the user. For the Jindalee OTHR, each track is given a confidence number expressed in the range $0, \ldots, 100$. For trackers that do not provide such a measure, a constant value can be recorded in the

TK data. The association process thresholds the tracks based on this confidence number, allowing the threshold to be varied, and potentially adapted to optimize the overall rating.

The tracker assessment process involves the following three steps:

1. For each source data set for assessing the tracking system, perform the following actions:
   (a) Generate the IT data.
   (b) Run the tracker being assessed on the data set.
   (c) Perform association and assessment using the IT and TK data.
2. Combine the results from the association and assessment on each data set
3. Rate the tracking system using predefined performance criteria

When alternative trackers are compared or when a repeated assessment of a tracker after parameter changes is done, there is no need to redo the IT data generation step given in 1(a). The rules for association and assessment are now given in more detail.

2.1 Association Rules

The association process cited in Figure 1 is now described with the aid of OTHR data features. This association process can be modified for other sensors with different measurement and tracking coordinates. The association rules use a normalised distance measure to compare TK and IT tracks, given by

$$d = \delta y^{\mathrm{T}} \mathbf{R}^{-1} \delta y, \tag{1}$$

where δy is the difference vector in the relevant coordinates, and $\mathbf{R}$ is a matrix of covariances based on the sensor measurement noise. For association with OTHR data, two distances are used. The first measures distance in range and azimuth and is denoted d_2. The second measures distance in range, azimuth and Doppler and is denoted d_3. Doppler must be treated specially for OTHR because the observed Doppler is ambiguous. The Doppler shift due to a moving object is proportional to its radial velocity. However, due to the relatively low waveform repetition frequency of OTHR, this frequency shift is often aliased and the apparent radial velocity determined by scaling the observed Doppler shift is different from the true radial velocity by a multiple of the ambiguous speed. The ambiguous speed can be easily determined from the waveform parameters. A track that has not correctly resolved the velocity ambiguity will be referred to as *misclassified.* In order for misclassified tracks to be associated with IT tracks, the d_3 distance uses the TK radial velocity shifted by that multiple of the ambiguous speed that minimises the Doppler component of δy.

The following rules are used to label each track as either *unassociated*, *associated* or *divergent* at each point in time.

A TK track is *unassociated* at time k if it is neither *associated* nor *divergent* (see below).

A TK track is *associated* at time k if it satisfies one of the following rules:

1. If the TK track was associated with an IT track at the previous time, $k-1$, and the 2D distance between the two is less than the divergence threshold, d_D, then the TK track remains associated with that IT track at time k.
2. If there is no association maintained from time $k-1$ (either because this is the first time point for the track, or the track was unassociated at $k-1$ or because the associated IT track no longer exists or has a corresponding $d_2 > d_D$) then the TK track is associated with the IT track corresponding to the smallest d_3, provided that this d_3 is less than the association threshold, d_A.

The association and divergence thresholds, d_A and d_D, are user defined parameters.

A TK track is *divergent* from an IT track at time k if it satisfies all of the following rules:

- The TK track was associated with or divergent from the IT track at time $k-1$.
- The 2D distance, d_2, between the TK track and the IT track is greater than the divergence threshold, d_D.
- There is no IT track with a 3D distance, d_3, less than the association threshold, d_A.

A TK track is *misclassified* at time k if it is associated with or divergent from an IT track, and the difference in the Doppler of the two is more than a threshold, d_m. If a TK track was misclassified at time $k-1$ and the corresponding IT track no longer exists, then it remains misclassified.

Association and divergence are illustrated with two parallel IT tracks in the range versus time display in Figure 2. The TK track is initially associated with the IT track at shorter range. It moves away from that IT track and then becomes divergent. After a period, it moves inside the association threshold of the longer range track and becomes associated with it. A TK track that is associated with one IT track and then another, is said to have *swapped*, as described in the next section.

Each TK track is associated with one IT track, at most, using these rules. More than one TK track may associate with a particular IT track. Similarly, each IT track is associated

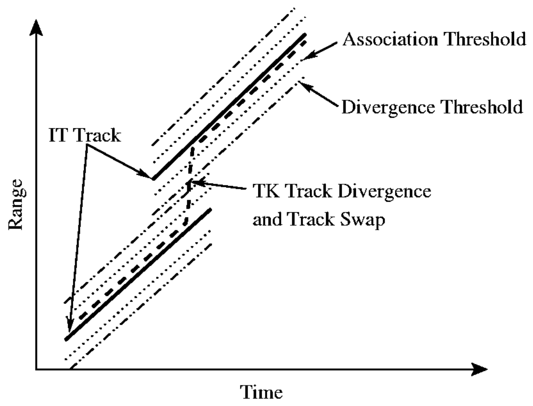


Figure 2: Track Association and Divergence

with at most one TK track. If the IT track was associated with a TK track at time $k-1$ and the TK track is still associated with the IT track (i.e. $d_2 < d_D$) then the association is maintained at time k. Otherwise the IT track is associated with the TK track with the lowest d_3 that is associated with it (i.e. $d_3 < d_A$). If there are no TK tracks associated with the IT track, then the IT track is unassociated at time k.

2.2 Categories and Metrics

The metrics identify the key imperfections that exist in tracking systems. They are grouped into four track categories, namely: establishment, maintenance, error and false tracks. The metrics are defined so that high values correspond to poor performance. The previous association rule parameters provide the data for evaluating the metrics.

2.2.1 Track Establishment Category

The metrics for this category measure the timeliness and reliability of the tracker to start valid tracks.

(i) *Track Establishment Delay:* Establishment delay is defined as the number of measurements from the start of an IT track to the first time that it is associated with a TK track. Establishment delay is quantified in terms of measurements rather than units of time (e.g. seconds) because it is the authors' experience that track maintenance rules tend to be expressed in terms of measurements, for example an M of N rule. Since the measurement rate may vary, especially for OTHR, time is not the best measure of delay. Note that a misclassified TK track can provide a valid association for determining establishment delay.

It is assumed that the IT track starts at the first detection from the corresponding target. It is also assumed that TK tracks are only recorded in the database once they have passed internal establishment tests, which will take at least a few measurements. This means that if a TK track associates with the IT track on the first point, then that TK track must have started on measurements not due to this target and then swapped onto the IT track. Thus any IT track with an establishment delay of zero is excluded from the analysis.

(ii) *Omitted tracks:* An omitted track is an IT track that has no associated TK track. The total number of omitted tracks is expressed as a percentage of the total number of IT tracks.

2.2.2 Track Maintenance Category

The metrics in this category measure how well a TK track maintains track on an IT track after establishment.

(i) *Track Overshoot:* This is the number of track updates that an associated TK track continues following the termination of the IT track without misclassification or association with another IT track. Overshoot is not measured in units of time (e.g.

seconds) for the same reason as establishment delay. Overshoot is not calculated for IT tracks that exit the surveillance region, or continue to the end of the data set.

(ii) *Missed Object History:* This is the percentage of time steps where an IT track has no associated TK track. The time period before the first association is not included, since this is already accounted for in establishment delay. The missed history is defined as the ratio of the number of time steps where there is no associated TK track to the number of time steps between the establishment of the first association and the termination of the IT track, expressed as a percentage. For the purposes of this metric, it does not matter which TK track is associated with the IT track at any point in time; track number changes are accounted for in other metrics. Note that a TK track with low confidence may give a poor missed history result because it becomes excluded from the analysis by the confidence threshold used during association at various times.

Missed history is not calculated for omitted tracks.

(iii) *Divergent Outbreaks*: A divergent outbreak is defined as the event of a TK track changing from associated to divergent. This metric counts the number of divergent outbreaks for each TK track.

(iv) *Divergence Length:* This length is given by the number of track updates that a track remains divergent after the track has been associated with an IT track. The time step count is from one at the start of TK track divergence. Divergence length is only counted for divergent tracks.

(v) *Track Swap:* This metric gives the number of times a TK track changes the IT track with which it is associated. Track Swap is measured for all associated TK tracks.

(vi) *Association Changes:* This is the number of changes in the TK track number used to report an IT track. For multiply associated TK tracks, the assumption is made that only one number, the original number, is reported while that track remains associated with the IT track. When this track stops and is replaced by another track, a track number change occurs. The number of association changes is measured for all IT tracks, except those that were omitted.

(vii) *Number Associated:* This is the total number of different TK tracks associated with an IT track. This will be at least one more than the number of association changes, and will be more if there are multiple TK tracks associated with the IT track at the same time. As above, it is not calculated for omitted tracks.

2.2.3 Track Error Category

For each associated TK track t, the track error is the sample standard deviation of the difference between TK track state $\hat{x}_t(k)$ and the associated IT track state $x_t(k)$ over the number of track updates where the object is not manoeuvring, K_t. This error is given by:

$$\hat{\mu}_t = \sqrt{\frac{\sum_{k=1}^{K_t}(\hat{x}_t(k) - x_t(k))^2}{K_t - 1}} \qquad (2)$$

A true measure of error can only be obtained using ground truth for recorded data or, for simulated data, the object parameters.

When IT tracks are derived from TK tracks produced from recorded data, an estimate of the error may be found by altering the states $x_t(k)$ in equation (2). This is done using a least squares fit of a straight line through the IT track points when the IT track is not manoeuvring. This defines an initial starting position and the velocity over the length of the track. To remove any offset between TK and IT tracks, their average position can be found for the relevant interval. The difference between these averages for the TK and IT tracks is added to the initial starting position for the least squares straight line.

The error metric is the sample standard deviation defined in equation (2) for the following coordinates:

1. *Track Position Error/Scatter*
2. *Track Speed and Heading Error/Scatter*

2.2.4 False Track Category

Misclassified tracks are grouped under this category. Thus, a misclassified track is considered to be a falsified track. The metrics in this category are defined in terms of a number or a corresponding length.

(i) *False Track Rate:* A false track is a TK track that is not associated with any IT track. The rate is expressed as the number per hour of data duration.

(ii) *False Track Length:* The length of each false track is measured in time steps.

(iii) *Misclassified Track:* Some trackers with multiple models may allow the track to switch between these models. This metric counts the number of track segments for each TK track over which the track is misclassified, ie has the wrong velocity ambiguity.

(iv) *Misclassified Track Length:* This is measured for each TK track that has a period of misclassification and is the number of time steps that the track is misclassified.

2.3 Rating and Overall Performance

The previous section defined a collection of metrics to assess tracker performance over a variety of areas. For most of these metrics, there is a value for each TK or IT track. This means that the assessment provides a sample distribution for each metric. While this information is useful, it may be desirable to provide a summary statistic for each metric that rates the tracker performance as a single number. Further, where a hard decision between trackers is to be made, it is desirable to provide an overall performance rating that summarises all of the information contained in the various metrics. Whereas the metrics are defined such that a low value indicates good performance[1], it is more intuitive

[1] This convention is necessary since some metrics, like establishment delay, are open ended.

to represent a good rating with a high score. Thus, the metric ratings and the overall rating will be expressed as a percentage score with 100% representing ideal performance.

The aim is to provide either relative or absolute performance. Relative performance is useful for comparing tracking filters, whereas absolute performance is useful for contract compliance. The measurement of absolute performance for contract compliance requires a large number of data sets over a range of object and clutter conditions. Also, there is a need for an error estimate of the measured performance. Usually tracker requirements for a contract only cover a subset of the metrics listed here and performance requirements are in terms of a specific metric such as the track establishment delay.

By contrast, tracker comparisons entail running different trackers on the same data, and then evaluating the difference in performance. The metrics define what is currently seen as the possible defects of a tracking system. For this reason they are more relevant to tracker comparison. This role is facilitated by being able to combine the metrics to give a single measure of performance. An approach that does this is described below. No attempt is made to deal with absolute performance.

The metrics measure performance in different units, e.g. time steps and number per hour. Also, some metrics are measured for each track while some are a single value over the data set. The metrics are assumed to come from a nonparametric distribution. That is, the metric values come from an unknown probability density function (pdf) which varies between trackers and has unknown parameters.

The approach taken for rating the trackers against the metrics is to define performance criteria for each metric and to determine the proportion of metric observations that meet the criteria. This may be viewed as an approximation to the probability that a particular instance of a specific metric will meet the relevant criterion. This process is now explained in more detail.

A sample cumulative distribution function (cdf), $F_m(s)$, is formed for each metric, where m is a metric index. Note that the domain of s over which each cdf, $F_m(s)$, is defined will be different since different metrics are defined over different units and intervals. Two sample points are chosen to summarise each cdf. These are referred to as the *baseline* and *goal* criteria and are denoted by s_m^b and s_m^g respectively. The baseline value in intended to represent an acceptable performance level, whereas the goal is intended to represent a preferred performance level. Since lower values of the metric represent better performance, the goal criterion should be lower than the baseline, i.e. $s_m^g < s_m^b$. The criteria represent user requirements. Although the setting of these values is subjective, it should be done *before* data are observed. Otherwise, the temptation is to adjust the criteria values in order to amplify the effect of a particular distribution feature on the overall rating. This damages the objectivity of the process and in the extreme provides a means for distorting the results in order to provide the answer expected by the user's prejudice.

The value of the cdf at each criteria defines the corresponding rating. For example, $F_m(s_m^g)$ is the goal rating for metric m. This value behaves in the desired fashion, where a value of $F_m(s_m^g)$ close to unity indicates that almost all of the observed tracks meet the goal criterion and a value close to zero indicates that almost none do. That is, higher ratings correspond to superior performance. This process is illustrated in Figure 3.

Note that the cdf illustrated in Figure 3 is for a metric with a discrete domain (e.g.

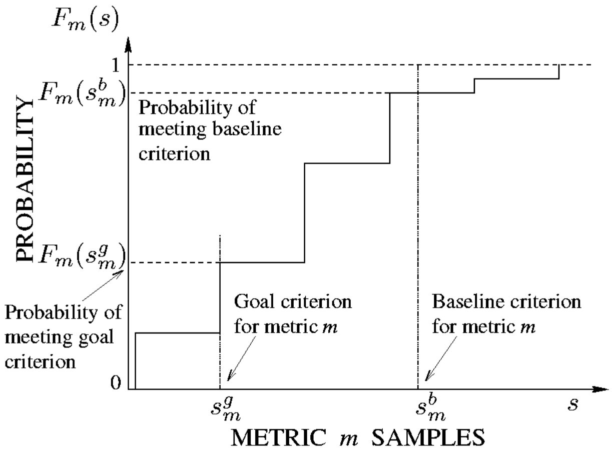


Figure 3: Sample cdf with example of baseline and goal criterion.

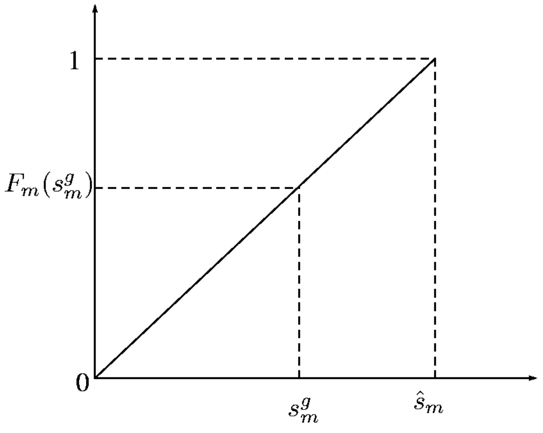


Figure 4: Rating for single value metrics.

establishment delay is measured in terms of the number of measurements). The staircase cdf is actually discontinous at these discrete points, and $F_m(s)$ is given by the greater value. This means that the rating is the proportion of observations that meet the criterion or do better.

Some metrics give only one value for a data set (the proportion of omitted tracks and the false track rate). In this case, it is not appropriate to form a cdf as above, since this will always give a rating of zero or 100%. If the observed metric value is less than the criterion, then a rating of unity is given. Otherwise, a line joining the origin with the observed metric value, $\hat{s}_m$, at a height of unity is constructed. The intersection of this line and a vertical at the criterion point gives the score, as shown in Figure 4. It is simple to show that the intersection point is given by $(s^g_m, s^g_m/\hat{s}_m)$ for the goal criterion, and similarly for the baseline.

The rating for single valued metrics is therefore defined as

$$F_m(s) = \begin{cases} 1 & \text{if } \hat{s}_m < s, \\ \frac{s}{\hat{s}_m} & \text{otherwise.} \end{cases} \tag{3}$$

Note that this heuristic rating does not have any statistical interpretation, however it behaves in manner consistent with the desired rating characteristics, and is simple to calculate.

Each of the metrics has now been condensed to a baseline and a goal rating. However, the stated purpose was to produce a single summary figure for each metric and for overall performance. To this end, the *mean* metric rating is defined as

$$F_m = \sqrt{F_m\left(s_m^g\right) F_m\left(s_m^b\right)}. \tag{4}$$

The individual metric performance values are then combined to give an overall performance by a weighted summation process. Namely,

$$\bar{F}_{goal} \equiv \sum_m w_m F_m(s_m^g), \tag{5}$$

$$\bar{F}_{base} \equiv \sum_m w_m F_m(s_m^b), \tag{6}$$

$$\bar{F}_{mean} \equiv \sum_m w_m F_m, \tag{7}$$

where w_m is the normalised weight for metric m. These normalised weights are determined from subjective importance scores. A category weight, W^c, is chosen for each of the four categories (establishment, maintenance, error and false tracks) reflecting the relative importance of each category. Then a metric weight is chosen for each metric within the category W_m^c. The normalised metric weight is then given by

$$w_m \equiv \frac{W^c W_m^c}{\sum_{j=1}^{4} W^j \sum_{k \in W^c} W_k^c}, \tag{8}$$

where c is the category to which metric m belongs, and $\sum_{k \in W^c}$ is a sum over all the metrics in category c. This process allows the subjective importance weights to be specified in an arbitrary range, and ensures proper normalisation. Any approach for defining weights is applicable, provided that $\sum_m w_m = 1$.

As discussed with the selection of the criteria, it is important that these weights are chosen before data are analysed. Altering the weights will change the relative influence of different metrics on the overall rating.

The weights used in this report were determined by polling a number of target tracking practitioners and averaging their responses. These values are shown in Table 1.

Table 1: Metric and category weights.

Metric Category	**Category Weight**	**Metric**	**Metric Weight**
Track Establishment	10	Establishment (updates)	10
		Omitted tracks (%)	7.4
Track Maintenance	7.2	Overshoot (updates)	7.8
		Missed object history (%)	9.4
		Divergent outbreaks	6.4
		Divergence length (updates)	6.4
		Track swaps	7
		Association changes	6.4
		No. Associated tracks	5.2
Track Error	3.8	Track Error - Range	7.6
		Track Error - Azimuth	7
		Track Error - Speed	6.6
		Track Error - Heading	6.4
False Tracks	6.8	False track rate	9.2
		False track length (updates)	7
		No. misclassified tracks	8.6
		Misclassified length (updates)	6.8

3 Variance of the Tracker Rating

The approach described in the previous section generates an overall rating for each tracker under test that is based on the aggregation of ratings for various assessment metrics. Each metric rating is derived from a finite amount of track data. This means that the metric ratings are random quantities, as is the overall rating. It is important to quantify the statistical variability in the tracker rating, otherwise it is impossible to discern whether the ratings of two alternative algorithms are significantly different in a statistical sense.

It would be desirable to derive confidence intervals for the tracker rating; however, this is difficult in practice. Instead, the variance of the tracker rating will be estimated. Since the metric and tracker ratings can only take values from zero to unity, their distributions must be compact, and are most likely asymmetric. This means that the common practice of translating variance into a confidence interval is not meaningful. For example, if for a mean of $\mu = 0.9$ and a variance of 0.01, a two standard deviation interval is $\mu \pm 2\sigma = 0.9 \pm 0.2 = [0.7, 1.1]$, which is outside the possible rating value range.

Most of the metric ratings are simply the fraction of tracks that meet the criterion for that metric. These are relatively easy to deal with and will be addressed first. There are two metric ratings that are calculated differently: the percentage of omitted tracks, and the false track rate. These are treated individually.

Once a variance is determined for each metric, an overall variance is obtained by combining them.

3.1 Properties of a Single Metric Rating

Most of the metrics produce a rating which is the fraction of observed tracks that meet the particular criterion. This metric rating is the observed relative frequency of success, and can be viewed as an estimate of the parameter of a binomial distribution. For clarity, the track establishment metric will be discussed, and the same arguments can be applied to each of the other metrics except as discussed.

Assume that each track is an independent realisation of an unknown pdf for establishment delay. For any pdf, there is a certain probability that any particular track will be established within the criterion time, denoted by θ. This is simply the integral of the pdf from zero to the criterion point. If the tracks are now classed as either pass or fail, then it is clear that the distribution of the number of passes is a binomial distribution with parameter θ. This parameter is actually the *true* rating that would be assigned to the tracker for the metric if the establishment delay pdf were known. Since it is not known, θ is estimated by the relative frequency of passes, $\hat{\theta} = \frac{N_p}{N}$, where N_p is the number of tracks which passed, and N is the total number of tracks.

Using a Bayesian strategy, a prior distribution for θ can be assumed, and then the posterior distribution for θ given the tracks observed can be obtained. If the prior for θ is assumed to be a beta distribution, then the posterior is also a beta distribution, and they are given by:

$$\begin{aligned} p(\theta) &= \beta(\theta; \alpha, \beta) && (9) \\ p(\theta | N_p, N) &= \beta(\theta; \alpha + N_p, \beta + N - N_p), && (10) \end{aligned}$$

where

$$\beta(\theta;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}. \tag{11}$$

If $\alpha = \beta = 1$ is chosen, then the prior, $p(\theta)$, is uniform, and the mode of the posterior, $p(\theta|N_p, N)$, is the relative frequency, $\frac{N_p}{N}$. Thus the metric rating can be viewed as the most likely true probability of success given the measurements, when the prior distribution is uniform. The variance of this distribution is given by

$$\begin{aligned} var(\theta|N_p,N) &= \frac{(\alpha+N_p)(\beta+N-N_p)}{(\alpha+\beta+N)^2(\alpha+\beta+N+1)} \\ &= \frac{(1+N_p)(1+N-N_p)}{(N+2)^2(N+3)}. \end{aligned} \tag{12}$$

As mentioned earlier, the distribution $p(\theta|N_p, N)$ is asymmetrical and compact on the interval $0\ldots1$, so care should be taken in the interpretation of the variance.

3.2 Omitted Tracks

As described in section 2.3, the rating for the omitted track metric is given by

$$F_m(s) = \begin{cases} 1 & \text{if } \hat{s}_m < s, \\ \frac{s}{\hat{s}_m} & \text{otherwise,} \end{cases}$$

where $\hat{s}_m$ is the observed fraction of omitted tracks, i.e. $\hat{s}_m = N_o/N$, when N_o tracks are omitted from a total of N IT tracks.

Again, the percentage of omitted tracks can be viewed as the result of a binomial process with an unknown parameter θ, N realisations and N_o failures. The observed omitted track rate is again the mode of the posterior for θ using the same approach as in the previous section. Thus the pdf for θ is given by $\beta(\theta; 1 + N_o, 1 + N - N_o)$.

The omitted track rating is implicitly dependent on the true probability of a track being omitted, θ, since the observed proportion of omitted tracks, $\hat{s}_m$, depends on it. This dependence is now recognised by denoting the rating as $F_m(s;\theta)$.

The omitted track rating is an ad hoc function that is not related to statistical measures of the posterior of the percentage of omitted tracks. In particular, it is biased. So, rather than measure its variability using the distribution's variance, we use the expected squared deviation from the rating, namely

$$V(\hat{\theta}) \equiv E\left\{\left(F(s,\theta) - F(s,\hat{\theta})\right)^2\right\}. \tag{13}$$

If $\hat{\theta} = 1$ then

$$\begin{aligned} V(\hat{\theta}) &= E\left\{(F(s,\theta)-1)^2\right\}, \\ &= E\left\{F(s,\theta)^2\right\} - 2E\left\{F(s,\theta)\right\} + 1. \end{aligned} \tag{14}$$

The expectations above are over the observed proportion of omitted tracks, for example

$$E\{F(s,\theta)\} = \int_0^s \beta(\theta; 1+N_o, 1+N-N_o)\,\mathrm{d}\theta + s\int_s^1 \frac{1}{\theta}\beta(\theta; 1+N_o, 1+N-N_o)\,\mathrm{d}\theta, \tag{15}$$

which can be expressed as

$$E\{F(s,\theta)\} = B(s; 1+N_o, 1+N-N_o) + \frac{s(N+1)}{N_o}\left[1 - B(s; N_o, 1+N-N_o)\right], \tag{16}$$

where the incomplete Beta function is defined as

$$B(x;\alpha,\beta) \equiv \int_0^x \beta(\theta;\alpha,\beta)\mathrm{d}\theta. \tag{17}$$

Algorithms for evaluating the incomplete Beta function are provided as part of standard high level mathematical programming environments.

An expression for $E\left\{F(s,\theta)^2\right\}$ can be obtained in a similar fashion. The full expression for V is quite long, and is not shown here. The details can be found in appendix A.

3.3 False Track Rate

The rating for false track rate is calculated in the same way as for the omitted tracks, except that the estimated parameter, $\hat{\theta}$, is the observed rate of false tracks. This quantity is the number of tracks per unit time, and cannot be viewed as an observation of a binomial process. Instead, it can be modelled as a Poisson arrival process with an unknown rate parameter. That is, the probability of observing N false tracks in a data set of duration T is given by

$$P(N) = \exp\{-\lambda T\}\frac{(\lambda T)^N}{N!}, \tag{18}$$

where λ is the true (unknown) rate of false tracks.

Similarly to the omitted tracks, the expected squared deviation is used to quantify the variability in the false track rate, and is defined in the same way. The expressions for the required statistical quantities are lengthy, and are given in appendix B.

3.4 Properties of the Overall Aggregate Rating

The aggregate rating is found by taking the weighted sum of the metric ratings. If the metrics were independent of each other, then the aggregate pdf would be the convolution of the metric pdfs, and the variance of the aggregate rating would be the weighted sum of the metric variances. However, the metrics are not independent; some of them are highly dependent on each other. For some metrics it is possible to estimate the dependence (or equivalently estimate the joint pdf) because they can be simultaneously observed. For example, each TK track that is associated with an IT track provides a measurement of: track error metrics, divergent outbreaks, divergence length, number of swaps, and misclassified track length. It is therefore possible to estimate the dependence between these metrics using the joint measurements. However, track overshoot, divergence length, and false track length are all measured in different ways. Intuitively, these ratings should be highly dependent because each measures the persistence of the tracker with various types of incorrect tracks. Since these metrics measure events that cannot be coincident, it is not possible to make any joint observation of the quantities. Thus it is not possible to infer the dependence between the ratings on these metrics.

In addition to the unobservable nature of the relationship between some metrics, other observable combinations depend highly on the baseline and goal criteria values. For example, if the goal criterion for the number of divergent outbreaks is zero, then a track can only fail the divergence length metric if it has already failed the divergent outbreaks metric. Conversely, if the goal criterion for the number of outbreaks were 10, then very few tracks (if any) would ever fail this metric, and the divergence length metric would be independent of the number of divergent outbreaks.

For the reasons described above, it is impractical to estimate the dependence between metrics. Further, even if the dependence were known, it would be difficult to analytically derive the aggregate rating pdf. This means that the pdf must be approximated. It is assumed that the aggregate rating is the mean of the aggregate pdf (although this will only be true in the limit for large data sets), and the statistical spread of the pdf can be quantified by the approximate variance. It is not possible to determine the true variance because the metric dependencies are not known.

The aggregate variance is given by

$$var(F_{agg}) = \sum_{m=1}^{M} w_m^2 var(F_m) + 2 \sum_{m=1}^{M} \sum_{n=1, n \neq m}^{M} w_m w_n cov(F_m, F_n) \tag{19}$$

where the cross term $cov(F_m, F_n)$ is defined as

$$cov(F_m, F_n) = E(F_m F_n) - E(F_m)E(F_n). \tag{20}$$

When the metrics are independent, then the covariance $cov(F_m, F_n)$ is zero. When two metrics are completely dependent, then it is given by $\sqrt{var(F_m)var(F_n)}$. In order to prevent presumptuous judgement, a conservative assumption for the covariance is used. Thus, it is assumed that all metrics are completely correlated, which maximises the resulting aggregate variance. Using this assumption, the aggregate variance can be written as

$$\sqrt{var(F_{agg})} \approx \sum_{m=1}^{M} w_m \sqrt{var(F_m)}, \tag{21}$$

i.e. the aggregate standard deviation is the weighted sum of the metric standard deviations. This provides a pessimistic variance estimate. The true variance will be less, but one can be assured that if the difference between tracker ratings is significant when compared with the pessimistic estimate, then it is truly significant.

4 Tracker Comparison

The assessment method described in the previous sections has been implemented in the Tracker Assessment Tool (TAT). The TAT software comprises two parts. The first is written in C and uses an X Windows Motif user interface. The second part is written in IDL and provides graphical displays for viewing assessment results. Graphical displays for the data are available, together with three functional shells: IT Editor, Association and Assessment. Track data and the processes for their association and assessment are shown in Figure 5. For convenience, the data from the tracking system are identified by a tracker identity trk_id. The TAT Association process produces a trk_id_date_time.AS file from the source TK and IT tracks, identified by a date and time then produces. IT and TK tracks are displayed in a geographic display (see Figure 6(a)) and range versus time display (see Figure 6(b)) format. An option is available to show the results of the TAT Association process on these displays as a line joining the IT and TK track positions at each time point. The TAT Assessment process uses the association file with the TK and IT tracks to compute the metric data for each track. These results are stored in a trk_id_date_time.RES file. The TAT results process computes and displays the sample cdfs for all metrics as shown in Figure 7(a). From this display the assessment process is invoked and the performance for each metric, and the overall performance for all metrics are derived after applying the metric weights. This display is shown in Figure 7(b). The rating variance is a new extension of the process and is not yet fully integrated into the display software. The values are currently written to the ratings output file, but are not shown on the graphical displays.

4.1 Forming IT Tracks from Recorded Data

The key to this assessment process is the formation of the IT track file. One approach to forming IT tracks from recorded data is now described for the case of OTHR data. Because the IT track data is in the same coordinate system and units as the TK track data, the IT data can be derived from the TK data output by a tracking system.

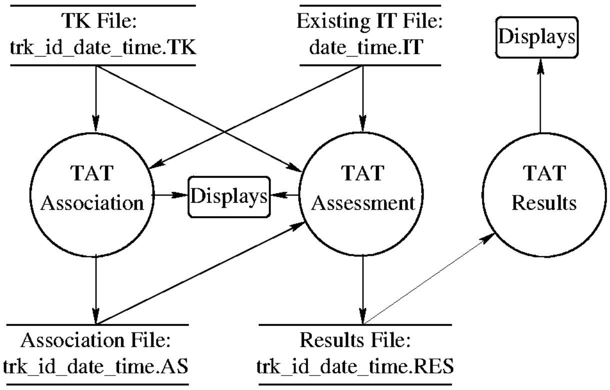
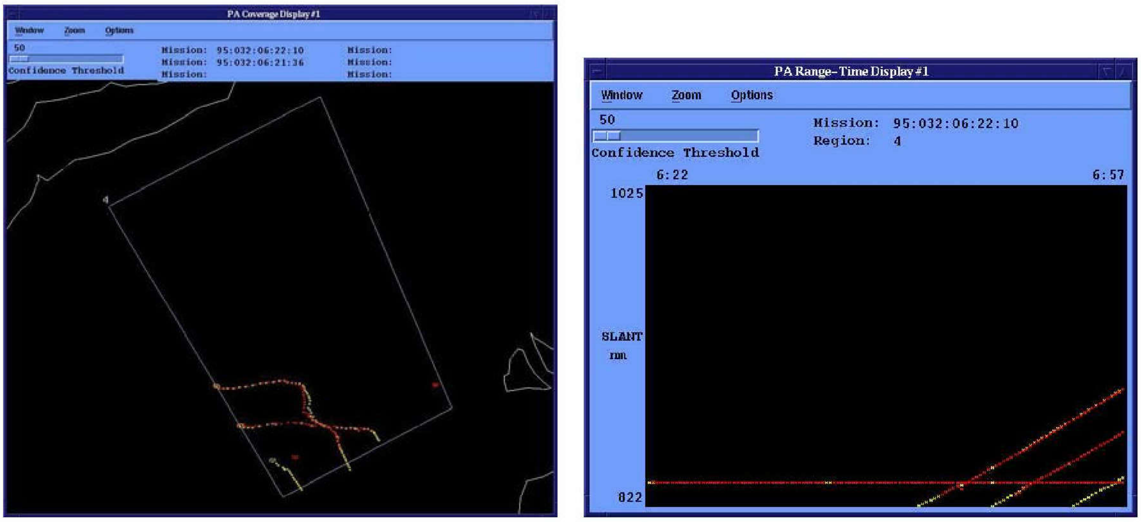


Figure 5: Organization of track data and processes for their assessment.



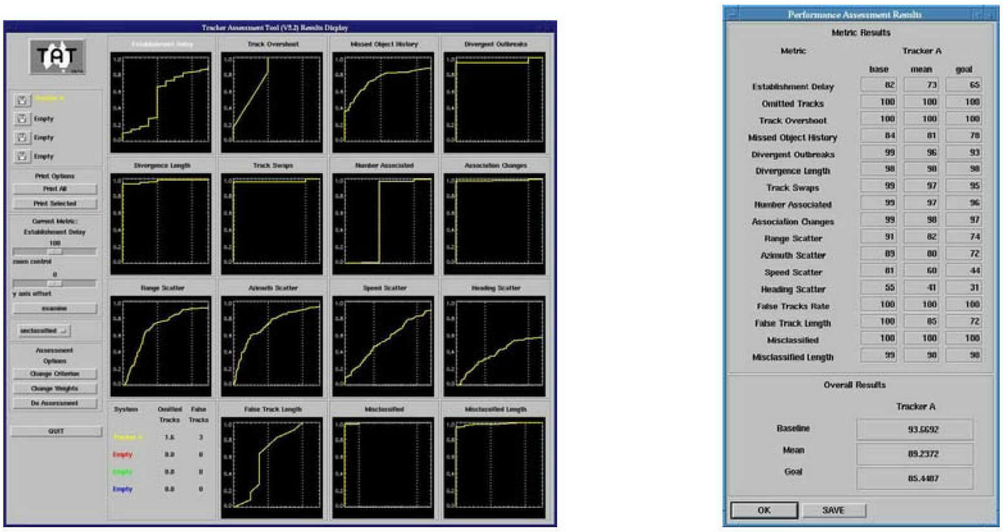
50
Confidence Threshold
Mission: 95:032:06:22:10
Mission: 95:032:06:21:36
Mission:
Mission:
Mission:
Mission:

50
Confidence Threshold
Mission: 95:032:06:22:10
Region: 4

(a) Coverage display

(b) Range versus time display.

Figure 6: TAT track data displays with TK and IT tracks



System	Omitted Tracks	False Tracks
Tracker A	1.6	3
Empty	0.0	0
Empty	0.0	0
Empty	0.0	0

Metric Results

Metric	Tracker A		
	base	mean	goal
Establishment Delay	82	73	65
Omitted Tracks	100	100	100
Track Overshoot	100	100	100
Missed Object History	84	81	78
Divergent Outbreaks	99	96	93
Divergence Length	98	98	98
Track Swaps	99	97	95
Number Associated	99	97	96
Association Changes	99	98	97
Range Scatter	91	82	74
Azimuth Scatter	89	80	72
Speed Scatter	81	60	44
Heading Scatter	55	41	31
False Tracks Rate	100	100	100
False Track Length	100	85	72
Misclassified	100	100	100
Misclassified Length	99	98	98

Overall Results

	Tracker A
Baseline	93.6692
Mean	89.2372
Goal	85.4487

(a) Cumulative Distributions

(b) Results Display from Metric Criteria.

Figure 7: TAT Metric Results Displays

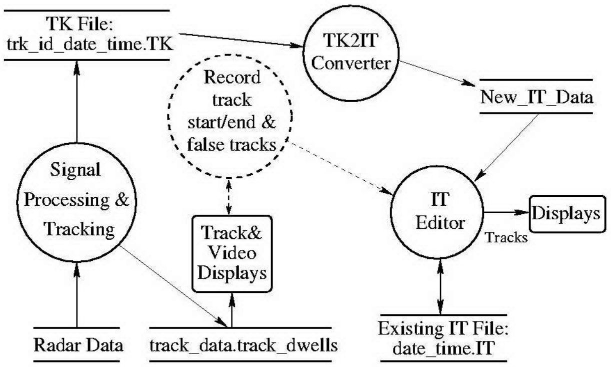


Figure 8: Steps to form an IT file.

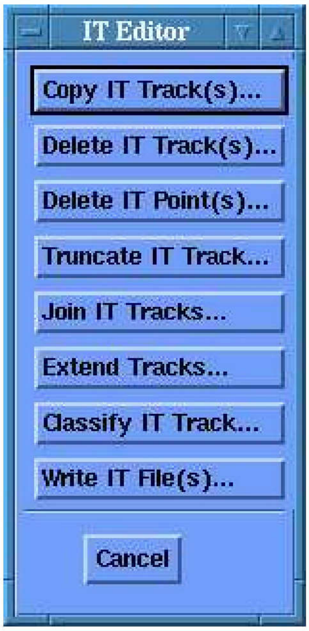


Figure 9: IT Editor Options

Figure 8 shows the data and processes used to form the IT track data. The TK file output from the Signal Processing and Tracking system is converted to a new IT file by the TK2IT process. Editing of this IT file (or an existing IT file) may be performed using the IT Editor. The IT Editor provides the options to copy, join, delete, truncate and project tracks forward and backward in time (see Figure 9). IT editing is performed based on information collected from replaying the data with a manual inspection of tracks overlayed on animated video data displays. This process supplies the track number for false tracks. For valid tracks, the track number is recorded with the start and end times from the underlying detection data.

The data replay allows manual track initiation to start tracks not detected by the tracking system. Where two tracking algorithms detect different targets that are all manually verified, then both TK files can be converted into IT files, and the IT editor can be used to copy the missed tracks from one IT file to the other.

The IT file format allows for a classification tag to be added to each IT track. Currently, this tag may be given the values: ‘standard’, ‘manoeuvring’, ‘decoupled’ or ‘fading’. The purpose of these tags is to allow the user to preclude certain types of track from the analysis. The standard and manoeuvring classes are self explanatory. The decoupled class is used for transponders and calibration signals that show an artificial Doppler shift even though they are stationary targets. The fading class was specially created for problematic targets. There are some targets that have very few detections, and some where it is difficult to decide whether there really is a target at all. For such targets, the tracker should not be penalised for failing to form a track (since the operator is uncertain). However, neither should the tracker be penalised if it does form a track. The solution is to have an IT track classified as fading. If there is a TK track, that associates with the fading IT track, it is not declared a false track. The TAT is instructed not to include fading tracks for any metrics.

In practice, the formation of the IT data from recorded data is an iterative process.

Even with these aids, achieving data integrity is a very time consuming manual exercise. The time taken increases with the complexity of the target conditions and the authors' experience is that it can take many hours to produce a reliable IT file for data sets of any significant length. For track assessment over some hundreds of data sets, this time becomes prohibitive and the process may be impractical. Nevertheless, recorded data does allow tracking performance assessment under realistic conditions that cannot be satisfactorily approximated by simulation.

4.2 Spawning IT Data

The process of forming IT data files is time consuming and prohibits the creation of a large library of assessment data. IT data spawning is one approach proposed to improve the efficiency of manual effort in the formation of IT files. The philosophy is to observe a region of interest at a high revisit rate (often referred to as stare mode operation), and then separate the data into interleaved data sets. The IT file is generated at the high sample rate, and then it is down-sampled to produce interleaved IT files. In this way multiple data sets are formed from a single IT gathering effort. This procedure was formulated during a USA and Australian cooperative programme in the late 90s to compare OTHR tracking systems [Allen et al. 1998]. The basic steps in this procedure are as follows:

1. Record data with a higher than usual measurement rate.
2. Replay the data and produce the IT file.
3. Form a down-sampled data set by selecting every nth time step from the input data and the IT file, starting from the first time step
4. Form a second down-sampled data set by repeating step 3, but starting from the second time step. Continue to form n down-sampled data sets

The above produces n data sets. Provided that the initial data was collected at n times the usual revisit rate, these are now typical sensor data. Although the targets obviously behave in the same way for each data set, the measurements are independent of each other, so the data sets can be viewed as n realisations of a particular scenario. The down-sampling could be done at different rates to produce partially correlated files. In the results presented here, there has been no data spawning.

4.3 Combining the Results

The TAT software described so far performs tracker assessment on a single data set: the input is a TK file and an IT file, the output is a results file. In practice, multiple data sets are combined to give results more representative of the 'average' tracker performance. The process for combining the results from several data sets is now described.

A separate results file is formed for each data set. These are combined into a single results file so that an overall measure of performance can be obtained. The combination either forms a single number for all the data or a separate measurement for each TK or IT track.

The metrics for Omitted Tracks and False Tracks result in a single number for all the data. For the overall percentage of omitted tracks, the data are combined by separately adding the number of omitted tracks and the number of IT tracks in each data set.

In the case of the false track metric, this is expressed as the number per hour. To account for possibly irregular time gaps in the data, the duration of the data is determined for each data set and these are summed over the combined data set. The duration of each data set is the product of the number of time steps and the duration of each time step. For an OTHR, the duration of each time step is the sum of the coherent integration time and the inter-dwell time.

The false track rate can then be found from the sum of the number of false tracks for each data set divided by the sum of the durations of each data set. One problem with this approach is that the clutter conditions vary between data sets. False track performance is typically tested in 'severe' clutter conditions, whereas all trackers might give equally good performance in 'benign' clutter conditions. If most of the data sets contain 'benign' data, then the difference in false track performance under difficult conditions will be suppressed by the quantity of less challenging data. In an attempt to normalize the clutter conditions, a data set clutter density weight, κ_d, where d is the data set index, is introduced. For the Jindalee OTHR, this weight is the ratio of the number of peak detections in the clutter area divided by the total number of peaks. This ratio is averaged over the length of the data set. The modified false track rate is then:

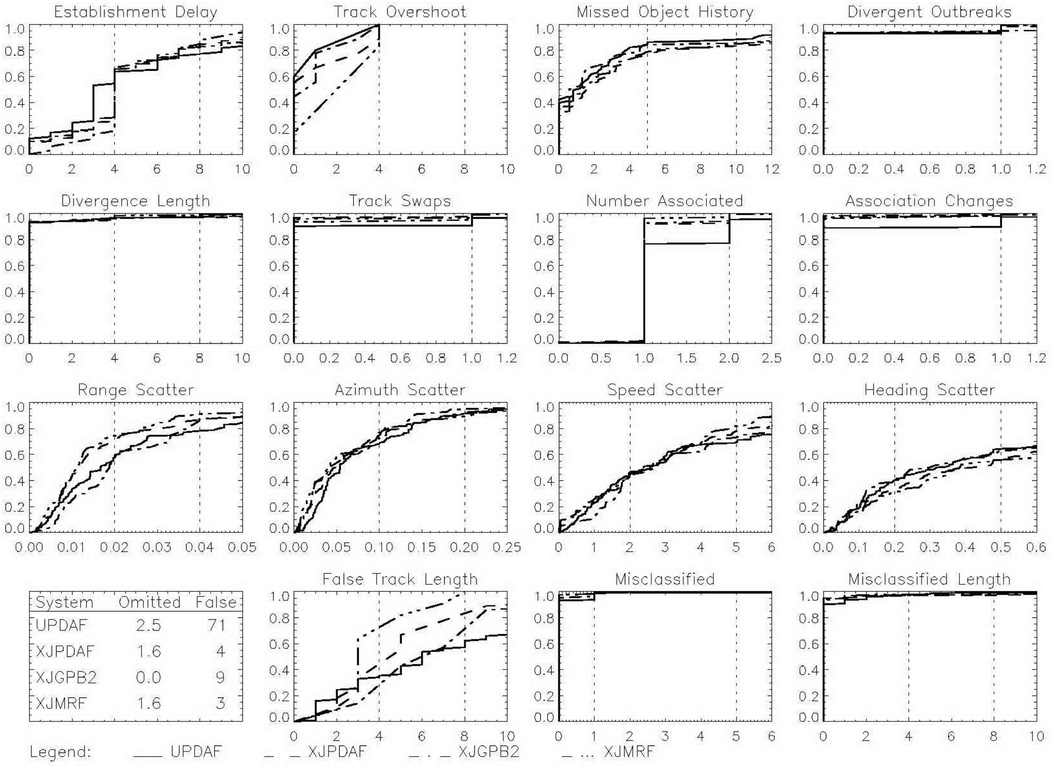
$$Adjusted\ False\ Track\ Rate = \frac{\sum_{d=1}^{D} \kappa_d\ No.\ False\ Tracks_d}{\sum_{d=1}^{D} \kappa_d\ Duration_d} \tag{22}$$

where D is the number of data sets to be combined. If all data sets have approximately the same weight, the result is similar to that without scaling. When there is a large variation in clutter density between data sets, the false tracks and duration for low clutter densities are scaled down so that the rate for severe clutter dominates the assessment for false tracks. This weighting approach is not applied to the false track duration metric.

All other metrics can give at least one value per TK track. Therefore, for these metrics, the combined data set is formed by successively appending the TK track data from each data set. No weighting is proposed as above for appending this track data.

4.4 Example Tracker Comparison

To illustrate the operation of the TAT, the results of assessing four tracking systems using recorded OTHR data are now presented. The main difference between the four tracking filters is the extent of past data retained after track update. The Probabilistic Data Association Filter (PDAF) approach represents the past data as a single mean and covariance for each object model. The PDAF filters that are assessed are the UPDAF [Colegrove 1999] and the XJPDAF [Colegrove et al. 2004]. The UPDAF is the current tracking algorithm run for the Alice Springs OTHR and has an adaptive clutter model, a multiple model filter for resolving radial velocity ambiguity, and a target visibility model for initiation and termination of tracks. The XJPDAF is an extended version that includes joint probability



System	Omitted	False
UPDAF	2.5	71
XJPDAF	1.6	4
XJGPB2	0.0	9
XJMRF	1.6	3

Figure 10: TAT Results

calculations, peak curvature [Colegrove & Cheung 2002] and an approximate exclusion model for to account for correlated measurements.

The third filter is a Generalized Pseudo-Bayes filter with 2 steps of past track history retained. It also has all the extensions developed for the PDAF and is referred to as the XJGPB2. The fourth filter has a variable length of retained history and is the Mixture Reduction filter. It has all the features in the XJPDAF included and is referred to as the XJMRF.

The test data consists of ten different data sets, containing mostly non-manoeuvring targets, although some manoeuvring targets are present. Some of the data sets contain interference sources that challenge the trackers' ability to suppress false tracks. The IT data is the result of an amalgamation of the output from the four tracking filters using the approach described in Section 4.1.

Figure 10 is the output from the TAT results display for the four trackers. This figure gives the sample cdfs for the metrics listed in Section 2.2 with the goal and baseline criteria shown by vertical dashed lines. There are some metrics that only have one line visible because the goal value coincides with the origin on the horizontal axis. The results display includes a table in the lower left corner that contains the percentage of omitted tracks and the adjusted false track rate.

These results illustrate the different distributions for each tracker's metric. The distributions show the metrics where each tracker differs in performance. The improvement from

Table 2: Performance Results

	UPDAF	XJPDAF	XJGPB2	XJMRF
Baseline	85 ± 3.5	93 ± 3.7	93 ± 5.0	94 ± 3.4
Mean	78 ± 3.9	88 ± 4.2	87 ± 5.7	89 ± 4.1
Goal	71 ± 4.1	84 ± 4.5	83 ± 6.0	85 ± 4.4

using the joint probability calculations in all but the MM-UPDAF is seen in the cdf for Number Associated and Association Changes. Noticeable differences are present in the false track rate and length and the percentage of omitted tracks. XJMRF has the lowest false track rate, which is 20 times lower than that of the UPDAF. The low false track rate makes it difficult to have a meaningful False Track Length cdf for the XJMRF. The TAT results display also provides the overall performance rating for the baseline and goal criteria with the geometric mean value. These are given in Table 2.

The table lists ratings to the nearest integer percentage with the tolerance number corresponding to one standard deviation. The variability increases from the baseline to the goal values because the goal ratings are lower, and are inherently more noisy. These results also show that the MM-XJMRF is the best performer for the data sets used in this test, although the differences between it, the MM-XJPDAF and the MM-XJGPB2 are well inside the tolerance levels. If the aggregate ratings are assumed to be Gaussian, then the probability that the MM-XJMRF is the best filter is approximately 40%. The MM-XJPDAF and the MM-XJGPB2 are each the best filter with a probability of approximately 30%. Thus the MM-XJMRF is most likely the best filter, but the evidence is far from conclusive. This difference is affected by the choice of baseline and goal criteria as well as the metric weights. If the criteria were set to higher values, then the differences between the trackers would be negligible.

5 Conclusions

A formalized procedure that rates and compares the performance of tracking systems has been developed and tested on recorded OTHR data. The sample distributions provided for each metric give a graphical assessment of the performance, which is useful for identifying performance deficiencies. However, these distributions depend on the data set and the target scenarios. Thus the results of a comparison only apply to the data set and target conditions used. The overall performance ratings summarise these results as a percentage. The method used to give the overall rating leads to the performance being dependent on the goal and baseline values as well as the metric weights. Thus, the rating between trackers depends on the data sets, object scenarios, metric weights, baseline and goal performance requirements. Because of this, the overall performance rating is primarily useful for comparing tracking systems on test data sets. The performance assessment procedure was demonstrated in an example application.

References

Allen, D., Colegrove, S. B., O'Neil, S., Arnold, J., Davey, S. J., Frank, V., Levine, P., Shaw, S. & Yssel, W. (1998) *Detecting and Tracking Targets with Over-the-Horizon Radars.*, Working Note WN 98B0000118, MOA.

Castella, F. R. (1986) Automatic track quality assessment in ADT systems, *in IEEE 1986 National Radar Conference Proceedings*, Los Angeles, California, USA, pp. 55–58.

Colegrove, S. B., Cheung, B. & Davey, S. J. (2003) Tracking system performance assessment, *in Proceedings of the 6th International Conference on Information Fusion*, Cairns, Australia.

Colegrove, S. B. & Cheung, B. (2002) A peak detector that picks more than peaks, *in Proceedings of Radar 2002*, Edinburgh, UK, pp. 167–171.

Colegrove, S. B., Davey, S. J. & Cheung, B. (2004) *The MM-XJPDAF: A JPDA Filter with Lots of Extras*, Technical report, under review, Defence Science and Technology Organisation, Australia.

Colegrove, S. B., Davis, L. M. & Davey, S. J. (1996) Performance assessment of tracking systems, *in Proceedings of the International Symposium on Signal Processing and its Applications*, Vol. 1, Gold Coast, Australia, pp. 188–192.

Colegrove, S. B. & Mabbs, S. A. (1991) Development and implementation of a performance assessment tool for radar signal processing and tracking, *in Proceedings of the 1991 TTCP Meeting of Subgroup K Technical Panel KTP-2.*

Colegrove, S. B. (1999) *Advanced Jindalee tracker: probabilistic data association multiple model initiation filter*, Technical report DSTO-TR-0659, Defence Science and Technology Organisation, Australia.

Mabbs, S. A. (1993) A performance assessment environment for radar signal processing and tracking algorithms, *Proceedings of the 1993 Pacific Rim Conference on Computers, Communications and Signal Processing* **37**(1), 214–225.

Mason, K. & O'Kane, P. A. (1992) Taxonomic performance evaluation for multitarget tracking systems, *IEEE transactions on Aerospace and Electronic Systems* **28**(3), 775–787.

Rothrock, R. L. & Drummond, O. E. (2000) Performance metrics for multiple-sensor, multiple–target tracking, *in Proceedings of the SPIE*, Vol. SPIE 4048, Orlando, Florida, USA, pp. 521–531.

(1994) *in Minutes of the 19th Meeting of Sub Group K Technical Panel KTP-2*, Vol. 1, DRA (Portsdown) and DGTE (West Freugh), UK.

Appendix A Variability of Omitted Track Rating

The omitted track rating is given by

$$F\left(s,\hat{\theta}\right) = \begin{cases} 1, & \text{if } \hat{\theta} \le s, \\ \frac{s}{\hat{\theta}}, & \text{otherwise,} \end{cases} \tag{A1}$$

where $\hat{\theta} = \frac{N_o}{N}$ is observed omitted track percentage, N_o is the number of omitted tracks, N is the total number of tracks, and s is the goal or baseline criterion.

The expected squared deviation of the omitted track rating is given by

$$V(\hat{\theta}) \equiv E\left\{\left(F(s,\theta) - F(s,\hat{\theta})\right)^2\right\}. \tag{A2}$$

If $\hat{\theta} = \le s$ then $F(s,\hat{\theta}) = 1$ and

$$\begin{aligned} V(\hat{\theta}) &= E\left\{(F(s,\theta)-1)^2\right\}, \\ &= E\left\{F(s,\theta)^2\right\} - 2E\left\{F(s,\theta)\right\} + 1. \end{aligned} \tag{A3}$$

The expectations above can be written in terms of incomplete Beta functions.

$$\begin{aligned} E\{F\left(s,\theta\right)\} &= \int_0^1 F(s,\theta)p(\theta)\mathrm{d}\theta && \text{(A4)} \\ &= \int_0^s \beta\left(\theta;\alpha,\beta\right)\mathrm{d}\theta + \int_s^1 \frac{s}{\theta}\beta\left(\theta;\alpha,\beta\right)\mathrm{d}\theta && \text{(A5)} \\ &= B(s;\alpha,\beta) + \int_s^1 \frac{s}{\theta}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}\left(1-\theta\right)^{\beta-1}\mathrm{d}\theta && \text{(A6)} \\ &= B(s;\alpha,\beta) + \\ &\quad s\frac{\Gamma(\alpha+\beta)\Gamma(\alpha-1)}{\Gamma(\alpha)\Gamma(\alpha+\beta-1)}\int_s^1 \frac{\Gamma(\alpha+\beta-1)}{\Gamma(\alpha-1)\Gamma(\beta)}\theta^{\alpha-2}\left(1-\theta\right)^{\beta-1}\mathrm{d}\theta. && \text{(A7)} \end{aligned}$$

A property of the Gamma function is that

$$\Gamma(x) = (x-1)\Gamma(x-1), \tag{A8}$$

so

$$\frac{\Gamma(\alpha+\beta)\Gamma(\alpha-1)}{\Gamma(\alpha)\Gamma(\alpha+\beta-1)} = \frac{\alpha+\beta-1}{\alpha-1}. \tag{A9}$$

Thus

$$\begin{aligned} E\{F\left(s,\theta\right)\} &= B(s;\alpha,\beta) + s\frac{\alpha+\beta-1}{\alpha-1}\int_s^1 \beta\left(\theta;\alpha-1,\beta\right)\mathrm{d}\theta && \text{(A10)} \\ &= B(s;\alpha,\beta) + s\frac{\alpha+\beta-1}{\alpha-1}\left[1 - \int_0^s \beta\left(\theta;\alpha-1,\beta\right)\mathrm{d}\theta\right] && \text{(A11)} \\ &= B(s;\alpha,\beta) + s\frac{\alpha+\beta-1}{\alpha-1}\left[1 - B(s;\alpha-1,\beta)\right]. && \text{(A12)} \end{aligned}$$

Substituting $\alpha = 1 + N_o$ and $\beta = 1 + N - N_o$ gives

$$E\{F(s,\theta)\} = B(s; 1 + N_o, 1 + N - N_o) + \frac{s(N+1)}{N_o}\left[1 - B(s; N_o, 1 + N - N_o)\right]. \quad \text{(A13)}$$

Using the same process, it can be shown that

$$E\{F(s,\theta)^2\} = B(s; 1 + N_o, 1 + N - N_o) + \frac{s^2 N(N+1)}{N_o(N_o-1)}\left[1 - B(s; N_o - 1, 1 + N - N_o)\right]. \quad \text{(A14)}$$

Thus $V(1)$ is given by

$$\begin{aligned} V(1) &= \frac{s^2 N(N+1)}{N_o(N_o-1)}\left[1 - B(s; N_o - 1, 1 + N - N_o)\right] \\ &\quad - \frac{2s(N+1)}{N_o}\left[1 - B(s; N_o, 1 + N - N_o)\right] \\ &\quad - B(s; 1 + N_o, 1 + N - N_o) + 1. \end{aligned} \quad \text{(A15)}$$

If $\hat{\theta} > s$ then $F(s,\hat{\theta}) = \frac{s}{\hat{\theta}}$ and

$$\begin{aligned} V(\hat{\theta}) &= E\left\{\left(F(s,\theta) - \frac{s}{\hat{\theta}}\right)^2\right\} \\ &= E\left\{F(s,\theta)^2\right\} - \frac{2s}{\hat{\theta}} E\left\{F(s,\theta)\right\} + \frac{s^2}{\hat{\theta}^2}. \end{aligned} \quad \text{(A16)}$$

Using the expressions above, this becomes

$$\begin{aligned} V(\hat{\theta}) &= \frac{s^2 N(N+1)}{N_o(N_o-1)}\left[1 - B(s; N_o - 1, 1 + N - N_o)\right] \\ &\quad - \frac{2s^2 N(N+1)}{N_o^2}\left[1 - B(s; N_o, 1 + N - N_o)\right] \\ &\quad + B(s; 1 + N_o, 1 + N - N_o)\left(1 - \frac{2sN}{N_o}\right) + \frac{s^2 N^2}{N_o^2}. \end{aligned} \quad \text{(A17)}$$

Appendix B Variability of False Track Rate Rating

As with the percentage of omitted tracks in Appendix A, the quantity of interest is

$$V(\hat{\theta}) \equiv E\left\{\left(F(s,\theta) - F(s,\hat{\theta})\right)^2\right\}, \tag{B1}$$

where θ is now the estimated false track rate N_F/T for N_F false tracks over duration T. As before, F takes either the value of unity or $\frac{s}{\theta}$ and the corresponding variability is

$$\begin{aligned} V(1) &= E\left\{F(s,\theta)^2\right\} - 2E\left\{F(s,\theta)\right\} + 1, \\ V(\hat{\theta}) &= E\left\{F(s,\theta)^2\right\} - \frac{2s}{\hat{\theta}}E\left\{F(s,\theta)\right\} + \frac{s^2}{\hat{\theta}^2}. \end{aligned} \tag{B2}$$

The expectations are given by

$$\begin{aligned} E\left\{F(s,\theta)\right\} &= \sum_{n=0}^{\infty} F\left(s, \frac{n}{T}\right) P(n) && \text{(B3)} \\ &= \sum_{n=0}^{\lfloor sT \rfloor} P(n) + \sum_{n=\lfloor sT \rfloor + 1}^{\infty} \frac{sT}{n} P(n), && \text{(B4)} \end{aligned}$$

and

$$\begin{aligned} E\left\{F(s,\theta)^2\right\} &= \sum_{n=0}^{\infty} F\left(s, \frac{n}{T}\right)^2 P(n) && \text{(B5)} \\ &= \sum_{n=0}^{\lfloor sT \rfloor} P(n) + \sum_{n=\lfloor sT \rfloor + 1}^{\infty} \frac{s^2T^2}{n^2} P(n), && \text{(B6)} \end{aligned}$$

where $P(n)$ is assumed to follow a Poisson distribution

$$P(n) = \exp(-\lambda T)\frac{(\lambda T)^n}{n!}, \tag{B7}$$

and λ is the true false track rate.

These expressions are calculated using the observed false track rate in place of the true λ.

DISTRIBUTION LIST

A Tracker Assessment Tool for Comparing Tracker Performance

S. B. Colegrove, S.J. Davey and B. Cheung

	Number of Copies
DEFENCE ORGANISATION	
Task Sponsor	
Director General Aerospace Development	1
S&T Program	
Chief Defence Scientist	1
FAS Science Policy	
AS Science Corporate Management	
Director General Science Policy Development	
Counsellor, Defence Science, London	Doc Data Sheet
Counsellor, Defence Science, Washington	Doc Data Sheet
Scientific Adviser to MRDC, Thailand	Doc Data Sheet
Scientific Adviser Joint	1
Navy Scientific Adviser	Doc Data Sheet and Dist List
Scientific Adviser, Army	Doc Data Sheet and Dist List
Air Force Scientific Adviser	Doc Data Sheet and Exec Summ
Scientific Adviser to the DMO M&A	Doc Data Sheet and Dist List
Scientific Adviser to the DMO ELL	Doc Data Sheet and Dist List
Platform Sciences Laboratory	
Director, PSL	Doc Data Sheet and Exec Summ
Information Sciences Laboratory	
Chief, ISRD	Doc Data Sheet and Dist List
Research Leader, WASB	Doc Data Sheet and Dist List
Head, TSF	1
S.B. Colegrove	1
S.J. Davey	1
B. Cheung	1
DSTO Library and Archives	
Library, Fishermans Bend	Doc Data Sheet

Library, Edinburgh	1 and Doc Data Sheet
Library, Sydney	Doc Data Sheet
Library, Stirling	Doc Data Sheet
Library, Canberra	Doc Data Sheet
Defence Archives	1
Capability Development Group	
Director General Maritime Development	Doc Data Sheet
Director General Capability and Plans	Doc Data Sheet
Assistant Secretary Investment Analysis	Doc Data Sheet
Director Capability Plans and Programming	Doc Data Sheet
Director Trials	Doc Data Sheet
Chief Information Officer Group	
Deputy Chief Information Officer	Doc Data Sheet
Director General Information Policy and Plans	Doc Data Sheet
AS Information Strategy and Futures	Doc Data Sheet
AS Information Architecture and Management	Doc Data Sheet
Director General Australian Defence Simulation Office	Doc Data Sheet
Director General Information Services	Doc Data Sheet
Strategy Group	
Director General Military Strategy	Doc Data Sheet
Director General Preparedness	Doc Data Sheet
Assistant Secretary Strategic Policy	Doc Data Sheet
Assistant Secretary Governance and Counter-Proliferation	Doc Data Sheet
Navy	
SO (SCIENCE), COMAUSNAVSURFGRP, NSW	Doc Data Sheet and Dist List
Director General Navy Capability, Performance and Plans, Navy Headquarters	Doc Data Sheet
Director General Navy Strategic Policy and Futures, Navy Headquarters	Doc Data Sheet
Deputy Director (Operations) Maritime Operational Analysis Centre, Building 89/90, Garden Island, Sydney	Doc Data Sheet and Dist List
Deputy Director (Analysis) Maritime Operational Analysis Centre, Building 89/90, Garden Island, Sydney	
Army	
ABCA National Standardisation Officer, Land Warfare Development Sector, Puckapunyal	Doc Data Sheet (pdf format)

SO (Science), Deployable Joint Force Headquarters (DJFHQ)(L), Enoggera QLD	Doc Data Sheet
SO (Science), Land Headquarters (LHQ), Victoria Barracks, NSW	Doc Data Sheet and Exec Summ

Air Force

SO (Science), Headquarters Air Combat Group, RAAF Base, Williamtown	Doc Data Sheet and Exec Summ

Joint Operations Command

Director General Joint Operations	Doc Data Sheet
Chief of Staff Headquarters Joint Operation Command	Doc Data Sheet
Commandant ADF Warfare Centre	Doc Data Sheet
Director General Strategic Logistics	Doc Data Sheet

Intelligence and Security Group

DGSTA, Defence Intelligence Organisation	1
Manager, Information Centre, Defence Intelligence Organisation	1 (pdf format)
Assistant Secretary Capability Provisioning	Doc Data Sheet
Assistant Secretary Capability and Systems	Doc Data Sheet
Assistant Secretary Corporate, Defence Imagery and Geospatial Organisation	Doc Data Sheet

Defence Materiel Organisation

Deputy CEO, DMO	Doc Data Sheet
Head Aerospace Systems Division	Doc Data Sheet
Head Maritime Systems Division	Doc Data Sheet
Chief Joint Logistics Command	Doc Data Sheet

Defence Libraries

Library Manager, DLS-Canberra	Doc Data Sheet

UNIVERSITIES AND COLLEGES

Australian Defence Force Academy Library	1
Head of Aerospace and Mechanical Engineering, ADFA	1
Deakin University Library, Serials Section (M List), Geelong, Vic	1
Hargrave Library, Monash University	Doc Data Sheet
Librarian, Flinders University	1

OTHER ORGANISATIONS

National Library of Australia	1
NASA (Canberra)	1

INTERNATIONAL DEFENCE INFORMATION CENTRES

US - Defense Technical Information Center	2
UK - Dstl Knowledge Services	2
Canada - Defence Research Directorate R&D Knowledge and Information Management (DRDKIM)	1 (pdf format)
NZ - Defence Information Centre	1

ABSTRACTING AND INFORMATION ORGANISATIONS

Library, Chemical Abstracts Reference Service	1
Engineering Societies Library, US	1
Materials Information, Cambridge Scientific Abstracts, US	1
Documents Librarian, The Center for Research Libraries, US	1

SPARES

DSTO Edinburgh Library	5

Total number of copies:	33

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA	1. CAVEAT/PRIVACY MARKING

2. TITLE A Tracker Assessment Tool for Comparing Tracker Performance	3. SECURITY CLASSIFICATION Document (U) Title (U) Abstract (U)
4. AUTHORS S. B. Colegrove, S.J. Davey and B. Cheung	5. CORPORATE AUTHOR Information Sciences Laboratory PO Box 1500 Edinburgh, South Australia, Australia 5111

6a. DSTO NUMBER DSTO–TR–1694	6b. AR NUMBER AR 013-351	6c. TYPE OF REPORT Technical Report	7. DOCUMENT DATE March 2005

8. FILE NUMBER 2005/1022387	9. TASK NUMBER JTW 03/148	10. SPONSOR DGAD	11. No OF PAGES 28	12. No OF REFS 12

13. URL OF ELECTRONIC VERSION http://www.dsto.defence.gov.au/corporate/reports/DSTO–TR–1694.pdf	14. RELEASE AUTHORITY Chief, Intelligence, Surveillance and Reconnaissance Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT *Approved For Public Release* OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SOUTH AUSTRALIA 5111
16. DELIBERATE ANNOUNCEMENT No Limitations
17. CITATION IN OTHER DOCUMENTS No Limitations
18. DEFTEST DESCRIPTORS Automatic tracking, Performance evaluation, Algorithms

19. ABSTRACT

This technical report describes a method for assessing the performance of automatic tracking systems against various metrics. These metrics are grouped in the categories of track establishment, track maintenance, track error and false tracks. A desired performance level is defined for each metric, and the probability that a particular track will meet this performance level is empirically approximated from archived sensor data. The individual metric probabilities are combined using a weighted average to produce an overall tracker rating. An important feature of the approach is that absolute truth is assumed to be unavailable, and output of the trackers under test is instead compared with *incomplete truth*. The incomplete truth is produced by a manual inspection of the data and through editing the output of the best available tracking algorithm. This report also discusses an implementation of the assessment method for Over-the-Horizon Radar. This software package is referred to as the Tracker Assessment Tool, and is used to compare four alternative tracking algorithms.